

Published in final edited form as:

Gene. 2007 December 30; 406(1-2): 1–7.

A basal deuterostome genome viewed as a natural experiment

R. Andrew Cameron and Eric H. Davidson

Biology Division 156-29, California Institute of Technology, Pasadena, California 91125, USA

Abstract

With the determination of its genome sequence the utility of the sea urchin model system increases. The phylogenetic position of the sea urchin among the deuterostomes allows for informative comparisons to vertebrate research models. A combined whole genome shotgun and bacterial artificial chromosome based strategy yielded a high quality draft genome sequence of 814 Mb. The predicted gene set estimated to include 23,300 genes was annotated and compared to those of other metazoan animals. Gene family expansions in the innate immune system are large and offer a first glimpse of how the long-lived sea urchin defends itself. The gene sets of the sea urchin place it firmly among the deuterostomes and indicate that various gene family -specific expansions and contractions characterize the evolution of animal genomes rather than the invention of new genes

Keywords

evolution; gene sets; gene homology

2. Introduction

Bilaterian animals that live in the ocean display an enormous amount of morphological and physiological variation. Their genomes embody the historical results of adaptation to specific living conditions that they and their ancestors have encountered over the more than one half billion years of existence on the planet. Thus the individual modern species that result from these various histories are “natural experiments” in that they reflect an accumulation of the successful properties on which natural selection has acted. There is subsequently a wide range of relatively simple biological systems in which particularly successful mechanisms of development and physiology are most easily examined. Furthermore, comparisons between the various animals emphasize the processes and mechanisms that have been generally most successful as well as those on which evolutionary change pivots. In the end, the light shed on particular biological processes in model systems more widely illuminates animals in general.

The gametes and embryos of the sea urchin have historically been such a model system for study of mechanistic developmental biology. The adults are easily available and the gametes are readily shed in a mature state. The embryo is relatively simple and optically clear thus experimentally tractable as well as visually beautiful. These features led to the choice of this model for molecular studies in developmental biology (Ernst, 1997; Pederson, 2006). The molecular studies on gene expression and cellular interaction that built upon the cell biology

Address for Correspondence R. Andrew Cameron, Beckman Institute 139-74, California Institute of Technology, 1200 East California Blvd., Pasadena, CA 91125, email: acameron@caltech.edu, phone: 626-395-8421, fax: 626-795-3382.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

and blastomere manipulation of the early 1900s have most recently led to the description of the gene regulatory networks that drive development (Davidson, 2006).

The modern milestone in the use of this model system is the determination of a draft genomic sequence for the purple sea urchin, *Strongylocentrotus purpuratus*. From a single wild individual, DNA and bacterial artificial chromosome libraries were used at the Baylor College of Medicine, Human Genome Sequencing Center (BCM-HGSC) to determine the genome sequence. A large number of newly completed expressed sequence tags from arrayed libraries in the Sea Urchin Genome Resource (<http://sugp.caltech.edu/>) were also determined. These sequence resources are the basis for a full predicted gene set for the sea urchin. From the annotation of almost half of the estimated 23,300 genes in the genome a new view of the sea urchin has emerged.

Here we discuss the first insights into the genome structure of the purple sea urchin and compare it to other genomes of metazoan animals. We detail the problems of sequencing a large, highly polymorphic genome and the solutions that accomplished the goal. We provide details of the experimental utility of the sea urchin model system and show where the genomic information is informing these experiments.

3. Phylogeny and genome characteristics of metazoans

There had been much discussion about the classification of the bilaterian animals proposed almost a century ago (Grobstein, 1908). The division of bilaterian phyla into deuterostomes and protostomes was flimsily based on a few developmental characters and it has become generally accepted only recently with the publication of molecular systematic studies. In the process the controversy about the assignment of previously enigmatic taxa to these major branches (Figure 1) has been set to rest. The division of the protostomes into ecdysozoans and lophotrochozoans has been an additional recent classification resulting from molecular assessments (Halanych et al., 1995; Aguinaldo et al., 1997). Now we can view the bilaterians as three large superphyla from which to make genome wide comparisons. The ecdysozoans are animals with a hard cuticle that is molted. They include the pan-arthropod groups (insects, crustaceans, spiders) and the nematodes as well as less well studied groups. In this superphylum the fruit fly *Drosophila melanogaster* and the nematode *Caenorhabditis elegans* are well studied developmental model systems. The lophotrochozoans are animals with either prominent ciliated adult feeding organs, the lophophore, or a characteristic larval form, the trochophore. They include the annelids, molluscs and their allies. There are no developmental models from the lophotrochozoans that has made the transition into the modern realm of molecular mechanistic developmental biology in the same way as ecdysozoans and deuterostomes, with the possible exception of the leech. However, there are several animals from this group whose genomes are being sequenced to a draft state. The marine gastropod or limpet, *Lottia gigantea*; the polychaete annelid, *Capitella* spp; the opsisthobranch gastropod *Aplysia californica* and the fresh-water leech are all the subjects of sequencing efforts (http://genome.jgi-psf.org/euk_home.html and <http://www.genome.gov/10002154>).

The third super-clade, the deuterostomes are the branch of the bilaterians that include the echinoderms as well as the vertebrates (Castresano et al., 1998; Turbeville et al., 1994; Wada and Satoh, 1994). From whole genome sequence collections a new classification among the deuterostomes has been inferred. Using 146 nuclear gene sequences among 14 deuterostomes and 24 other slow evolving groups the tunicates are placed nearer the vertebrates than the cephalochordates (Figure 1; Delsuc et al., 2006). With the acquisition of sequences in these same genes for *Xenoturbella*, the hemichordates and echinoderms placed as the most basal group and the cephalochordates, the most basal chordates (Bourlat et al., 2006). This recent

classification sets the stage for useful genome comparisons among the deuterostome phyla and to a real extent the genomic evolution in this superclade.

The number of large bilaterian animal genomes that have been sequenced total about 40 in 2006. Many of these are animals with genome sizes near the smaller end of the range measured in this group (Figure 1). This bias is a practical one resulting from the need to efficiently use sequencing resources. Tunicates have reduced gene number and more compact genomes compared to other deuterostomes. Both *Oikopleura dioica* and *Ciona intestinalis* are similar in this regard indicating that it is likely shared by all of the urochordates. The larger gene number and genome size of the sea urchin whose phylogenetic position lies basal to the deuterostomes suggests that this is an independent derived feature of the tunicates. The larger genome size of the vertebrates is probably due to two whole genome duplications leading to the vertebrates and even more than two among the fishes (Ohno, 1970; Garcia-Fernandez and Holland, 1994). In these terms the sea urchin genome is a good comparison to examine the evolution of genes and genomes in the deuterostomes and possibly the bilaterians more generally (Materna et al., 2006).

4. Sequencing the genome of a highly polymorphic animal

The purple sea urchin genome is about 814 Mb as estimated from the sequencing project (Sea Urchin Sequencing Consortium, 2006). This is in close agreement with the 800 Mb \pm 5% previously determined biochemically (Hinegardner, 1974). A unique feature of the sea urchin genome in comparison to other large animal genomes that have been sequenced is the high degree of polymorphism or intraspecific sequence variation. Solution hybridization studies measured the polymorphism at 4–5% (Britten et al., 1978). The polymorphism remains the same whether two animals living in close proximity are measured or two animals from the opposite ends of their range. The lack of differences over distance implies that there is gene flow over the entire range. The latter conclusion has since been verified by molecular genetic studies (Palumbi and Wilson, 1987). Furthermore, the differences between genome sequences are due in part to insertions and deletions as well as single nucleotide polymorphisms (Britten et al., 2003). The mechanism of this variation is not understood but the ancient origin and large founding population probably contributed to this quality.

The high degree of polymorphism called for new approaches to determining the genome sequence. It is expected that as the sequence coverage approached 20X, two genomes would be assembled separately from an individual DNA sample. Unfortunately such extensive sequencing of a large animal genome is beyond the availability of resources at present. The HGSC-BCM employed a combined strategy where whole genome shotgun reads to 6X coverage were assembled with 2X coverage of a BAC-based minimum tiling path of 8248 BAC clones (Sea Urchin Sequencing Consortium, 2006; Sodergren et al., 2006). The clone-array pooled shotgun method (CAPSS; Cai et al., 2001) was used to reduce the number of sequencing libraries needed to sequence the BACs. In this method pools of BAC clones in an array were sequenced together and then deconvolved computationally. Components were added to the Atlas assembler in order to handle local regions of heterozygosity and thus take full advantage of data from the BACs, each of which is a single haplotype. Finally, a high-quality draft assembly (Spur_v2.1) with only 4–5% redundancy and more than 90% of the genome sequenced was obtained from only 8X coverage of the highly polymorphic genome. The sequence was contained in 54,960 scaffolds assembled from 105,692 contigs. The aggregate of the longest scaffolds that include 50% of the base pairs number 13,575 and the smallest among these is 142 Kb (N50). Furthermore, the completeness of the assembly was judged to be high since 95% of the ESTs from this species available in dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/>) gave matches to the genome sequence. Since future species to

be sequenced are likely to share this high degree of polymorphism, this strategy is a paradigm for future projects.

5. Annotation and gene model characterization

The assembled sequence was used to compute predicted gene models using four different algorithms: Genscan (Burge and Karlin, 1997; Guigo et al., 2000), FgenesH (Salamov and Solovyev, 2000; Solovyev, 2001), Ensembl (Potter et al., 2004) and the Gnomon prediction method from NCBI (Souvorov et al., 2004). These sets of predictions were combined using an algorithm nick-named “Glean” which is a Latent Class Analysis approach that assesses accuracy and error rates for each source of evidence and then assembles a consensus prediction set, the official gene set or *OGS* (Elsik et al., 2006). The total number of predicted gene models in the OGS from the first WGS assembly was 28,944. The several prediction programs and the Glean-derived OGS took advantage of the approximately 130,000 ESTs in Genbank and produced by the Baylor HGSC Genbank to strengthen the predictions. In addition, a whole genome tiling array was hybridized to mRNA from multiple early embryonic stages and the resulting positive tiles aided in the certification of predicted gene models while providing 3-prime untranslated sequences that are poorly predicted by the computational methods (Samanta et al, 2006). More than 240 investigators from the research community manually annotated over 10,000 of these predicted gene models.

Computational comparisons of the sea urchin gene predictions to other sequenced genome gene sets provide a rough overview of relationships between model systems and a first glimpse of the shared and unique features of these gene catalogs. To a first approximation a classification of gene models as to conserved domain reveals homology and possible function for the encoded proteins. Using common databases of conserved domains like PFAM (Sonnhammer et al., 1998) and Interproscan (Interpro Consortium, 2001) about 70 % of the Glean protein models could be assigned to at least one of 4182 different domains. Similar searches with the non-redundant protein sets for mouse, fruit fly, and nematode revealed the classes of proteins that are unique to sea urchin and the most common shared gene families (Materna et al., 2006). Two of the most striking expansions are the toll-interleukin1-receptors (TIR: PF01582) and the speract/scavenger receptors (SRCR; PF00530) (Table 1.). These families of genes are involved in innate immunity (discussed below). Domains that characterize proteins commonly associated with cell death processes are also more abundant than the families in the other model organism (Robertson et al., 2006). The PFAM motif for histones (PF00125) is contained in 335 predicted proteins in the sea urchin compared to 20 in the mouse and only 11 in the fly. Manual annotation of the sea urchin gene models containing histone genes demonstrates that there are three classes of these genes and the sea urchin ones are the most complex set yet encountered (Marzluff et al, 2006). This distribution suggests that the expansion of histone genes is likely to be an invention of the deuterostomes. Proteins identified by the hyaline motif (HYR; PF02494) make up a large class in the sea urchin. The HYR motif was originally identified in the sea urchin as an extracellular matrix protein involved in fertilization (Wessell et al, 1998). More recently it has been found in bilaterians generally associated with cell adhesion motifs (Callebaut et al, 2000). Three classes of zinc finger motifs identify almost 900 predicted proteins. Because these motifs are usually present in multiple forms in a protein, this number may be an over-estimate. Zinc finger structures are known to function in DNA binding and protein interaction. This may be an expanded group of proteins in sea urchins (Materna et al., 2006).

There are 1375 domains that identify proteins in other model organisms but are not found in sea urchins. Prominent among these are a class of proteins involved in oocyte maturation called spindlins, a unique class of C₂H₂ zinc finger transcription factors that bear a transcriptional repression domain called a Krab domain and the immunoglobulins of adaptive immunity

(Materna et al., 2006). The distinctive olfactory domains individually conserved in the genomes of sequenced model organisms are not present in the sea urchin even though the general class of rhodopsin-type G-protein coupled receptors (GPCR) with a 7-transmembrane structure is the most abundant one. However genomic signatures have been found for a sea urchin specific group of these receptors that suggest they may be olfactory in function (Raible et al, 2006).

6. An overview of gene homology

A particularly conservative method of identifying possible orthologs between two gene sets is the reciprocal best BLAST method that is particularly suited for broad comparisons of whole gene sets. (for example see Stuart et al., 2003). In this method, the first species sequence finds a target in the second species which in turn recovers the first sequence from that species set. If we compare among four deuterostome non-redundant gene sets (Figure 2A) the sea urchin has a greater number of matches to the two vertebrates than does the urochordate. This is to be expected since the urochordate genome shows signs of sequence loss or compaction and a smaller gene set than either the echinoderm or the vertebrates. The distribution of matches among the human, cnidarian, fruit-fly and sea urchin reflects the metazoan relationships (Figure 2B). In this distribution the fruitfly has fewer matches to the cnidarian and the sea urchin than the sea urchin has to the cnidarian. This result could be due to the condition where the ecdysozoan has a group of genes that are too highly diverged to cluster or has suffered some loss in gene diversity. When the sea urchin is compared to the other ecdysozoan gene set (nematode) the matches are about equal to those between the fly and the nematode (Figure 2C). This suggests that the loss or divergence of genes occurred independently in the two ecdysozoan phyla.

7. The sea urchin gene set defines defines the bilaterian gene set

The gene assignments by protein domain match and homology analyses give a clearer overview of the characteristics of the deuterostome gene catalog than can be obtained without the sea urchin data. The metazoan gene set is larger than expected from human and ecdysozoans, and the sea urchin gene set in concert with the cnidarian indicates reduction in gene complexity in ecdysozoans. On the other hand, the bilaterian gene set is smaller than expected from human and a small sample of ecdysozoans. From the perspective of the sea urchin gene set it now seems that the vertebrates invented fewer gene families than previously thought. In summary the sea urchin example shows that specific gene family expansions and contractions characterize new phylogenetic groups among the bilaterians rather than the abrupt invention of new gene classes.

8. Genomics and gene regulatory networks

Significant progress has been made in the study of the gene regulatory relationships hardwired in the genome, that control the development of early embryonic patterns and structures in sea urchin development (Davidson, 2006). The easily available embryos, their cellular simplicity and optical clarity contribute to the desirability of this organism as a model for these studies. But the most remarkable property is the capacity of the sea urchin zygote to efficiently incorporate endogenous DNA molecules (Flytzanis et al., 1988). After concatenation and amplification, these artificial DNA constructs are expressed in an identical manner to the genomic sequences from which they are derived. This gene transfer property of sea urchin zygotes has permitted the analysis of cis-regulatory interactions and the construction of gene regulatory circuits (GRN) that control the specification events in early embryogenesis (Oliveri and Davidson, 2004). From a set of operational definitions of early cellular processes manifest through maternal information and cellular communications, the regulatory interaction at the level of genomic sequence of cis-regulatory modules has emerged very clearly (<http://supg.caltech.edu/endomes/>). In many cases the site of binding of transcription factors that are

the ultimate nodes of these networks are characterized. As the elements in these networks increase in number, network motifs and sub-circuits are evident. These building blocks contain members that connect in well-defined and typical ways and each sub-element performs a discrete task (Ben-Tabou de-Leon and Davidson, 2006). Thus the ability to classify the dizzying complexity of these embryonic interactions offers the promise of higher order organization possible from this systems biology approach.

A draft genome sequence from a combined WGS and BAC strategy really facilitates the description of gene regulatory networks. An annotated set of predicted gene models simplifies the discovery of genes in networks. For example, the non-zinc finger transcription factors are fully annotated in the sea urchin genome (Howard-Ashby et al., 2006a, b) and their expression patterns have been described. Candidate members of a GRN are thus already limited by spatial expression pattern in the tissue territory of interest. The genomic sequence surrounding a candidate gene is available in a BAC sequence and the clone can be used for recombinant BAC studies.

9. Sea Urchin Immunity

The surprising longevity of sea urchins frames another class of natural experiments. From tagging studies the purple sea urchin has a maximum life span of about 50 yr. Its congener, the red sea urchin *Strongylocentrotus franciscanus*, may live more than 200 yr. Given this sort of longevity sea urchins would be expected to possess well developed systems for immunity and chemical defense. Indeed there is a huge expansion in the germ line repertoire of the genes that show clear homology to genes that function in innate immunity. These include 222 Toll-like receptor genes (TIR); 203 NACHT domain-LRR (NLR) genes; a large family of secreted response genes called 185/333 genes and 218 gene models encoding members of the superfamily of scavenger receptor-cysteine rich proteins (SRCR) (Rast et al., 2006; Hibino et al., 2006; Sea Urchin Genome Sequencing Consortium, 2006; Nair et al., 2005). Fully 4–5% of the genes in the predicted gene set are involved in immune function (Rast et al, 2006). The largest class of TIR genes are vertebrate like in structure and appear to have been duplicated recently. The expansion of this family coupled with the large number of pseudogenes in this class leads to the conclusion that this is a dynamically evolving set of genes (Hibino et al, 2006).

The identification of this remarkable set of genes in sea urchins offers a new perspective on immune function in this long-lived animal and belies the simplicity suggested by its unremarkable appearance. The sea urchin mounts a complex innate immune response from a wide array of recognition molecules. Due to their sheer abundance and diversity these immune effectors must contain previously unknown new elements of biological structure that recognize pathogen-associated molecular patterns the exploration of which may have great general utility.

10. Paleogenomics

Sea urchins display longevity in another dimension, too. The characteristic internal skeleton of echinoderms, the phylum to which sea urchins belong is the stereom (Bottjer et al., 2006). It first appeared in the fossil record just before the beginning of the Cambrian ~ 542 million years ago (Mya). The stereom is recognizable by its distinctive meshlike structure composed of calcium carbonate with a minor proportion (5%) of magnesium carbonate. The stability of this high-magnesium mineral has led to an abundant and well understood fossil record for echinoderms.

Because the skeleton is accessible in the spicule of the embryo, skeletal development is well studied. The specification of the cells that form the spicule, the primary mesenchyme cells, is described as part of the endomesoderm gene regulatory network (Davidson, 2006; Oliveri and

Davidson, 2004). The proteins that make up the matrix into which the mineral is deposited determine its crystalline form. Some of these proteins have been identified and one class exists in the genome as a large family of duplicated genes that are distinguished by a c-type lectin motif and a unique form of amino acid repeat. These proteins are dynamically expressed with different members of the family evident at different developmental stages (Livingston et al., 2006). Given that an ancient and unique skeletal element can be traced from the Cambrian to the present in the fossil record and that the element is constructed using a unique set of proteins suggests that these proteins are an ancient invention. Such a paleogenomics perspective lends a dimension of deep time to genomic studies (Bottjer et al., 2006).

11. Conclusions

The effort to sequence the sea urchin genome is itself an experiment in the broad sense. The sequencing strategy used to successfully overcome the extremely high polymorphism in the sea urchin genome represents a test case that will inform the sequencing of many of the larger animal genomes sure to follow. These prospective projects are concerned with genomes on the order of 500 Mb or larger. Furthermore, the best material to be had will contain two haplotypes in the DNA of a single individual, since isogenic strains and genetic maps are seldom found among these forms.

Information emerging about the sea urchin genetic toolkit serves as a very informative outgroup comparison to define the extent of the deuterostome characters at the genomic level and concomitantly what constitutes a bilaterian or a metazoan. In the broadest view many different processes shaped the final genetic content of animal genomes. There is no trend toward uniform expansion of gene families to parallel the increased complexity in the vertebrates for example. Unique expansions and losses characterize each taxonomic group. In summary evolution crafted these complex genomes through the dynamic changes in size of various gene families without the invention of many new elements.

The full promise of the genome content of the sea urchin continues to be realized. From gene regulatory networks to immune function questions lie ready for study and surprises await discovery. For the sea urchin is still an enigma that appears much less like human beings visually than its genome shows it to be.

Acknowledgements

We thank Emanuelle Morin and Kris Khamvongsa for technical assistance during this project. We also wish to acknowledge Dan Rokhsar, Joint Genome Institute, DOE for permission to use the unpublished gene models from the star anemone genome. This work was supported by the NIH RR15044, NSF IOB-0212869 and the Beckman Institute.

Abbreviations

BAC	Bacterial artificial chromosome
BCM-HGSC	Baylor College of Medicine- Human genome Sequencing Center
CAPSS	Clone array pooled shotgun strategy
EST	Expressed sequence tag
dbEST	EST database

GPCR	G-protein coupled receptor
GRN	Gene regulatory network
HYR	hyalin repeat protein
LRR	leucine-rich repeat protein
Mya	Million years ago
N50	Sequence assembly metric
NLR	Nacht-LRR protein
OGS	Official gene set
PFAM	Protein family database
SRCR	Scavenger receptor-cysteine rich
TIR	Toll interferon-like receptor
WGS	whole genome shotgun

References

- Aguinaldo AMA, Turbeville JM, Linford LS, Rivera JR, Garey JR, Raff RA, Lake JA. Evidence for a clade of nematodes, arthropods and other molting animals. *Nature* 1997;387:489–493. [PubMed: 9168109]
- Ben-Tabou de-Leon S, Davidson EH. Deciphering the underlying mechanism of specification and differentiation: The sea urchin gene regulatory network. *Sci STKE* 2006;(361):pe47.10.1126/stke.3612006pe47 [PubMed: 17106076]
- Bottjer DJ, Davidson EH, Peterson KJ, Cameron RA. Paleogenomics of echinoderms. *Science* 2006;314:956–960. [PubMed: 17095693]
- Bourlat SJ, Juliusdottir T, Lowe CJ, Freeman R, Aronowicz J, Kirschner M, Lander ES, Thorndyke M, Nakano H, Kohn AB, Heyland A, Moroz LL, Copley RR, Telford MJ. Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *NATURE* 2006;444:85–88. [PubMed: 17051155]
- Britten RJ, Cetta A, Davidson EH. The single copy sequence polymorphism of the sea urchin *Strongylocentrotus purpuratus*. *Cell* 1978;15:1175. [PubMed: 728997]
- Britten RJ, Rowen L, Williams J, Cameron RA. Majority of divergence between closely related DNA samples is due to indels. *Proc Natl Acad Sci USA* 2003;100:4661–4665. [PubMed: 12672966]
- Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 1997;268:78–94. [PubMed: 9149143]

- Cai WW, Chen R, Gibbs RA, Bradley A. A Clone-Array Pooled Shotgun Strategy for Sequencing Large Genomes. *Genome Research* 2001;11:1619–1623. [PubMed: 11591638]
- Callebaut I, Gilges D, Vigon I, Mornon J-P. HYR, an extracellular module involved in cellular adhesion and related to the immunoglobulin-like fold. *Protein Science* 2000;9:1382–1390. [PubMed: 10933504]
- Castresana J, Feldmaier-Fuchs G, Yokobori SI, Satoh N, Pääbo S. The mitochondrial genome of the hemichordate *Balanoglossus carnosus* and the evolution of deuterostome mitochondria. *Genetics* 1998;150:1115–1123. [PubMed: 9799263]
- Davidson, EH. *Gene Regulatory Networks in Development and Evolution*. Academic Press/Elsevier; San Diego: 2006. *The Regulatory Genome*.
- Delsuc F, Brinkmann H, Chourrout D, Philippe H. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature* 2006;439(23):965–968. [PubMed: 16495997]
- Elsik CG, Worley KC, Zhang L, Milshina NV, Jiang H, Reese JT, Childs KL, Venkatraman A, Dickens CM, Weinstock GM, Gibbs RA. Community annotation: Procedures, protocols, and supporting tools. *Genome Res* 2006;16:1329–1333. [PubMed: 17065605]
- Ernst SG. A century of sea urchin development. *Am Zool* 1997;37:250–259.
- Flytzanis, CN.; Hough-Evans, BR.; Britten, RJ.; Davidson, EH. Gene transfer by microinjection into the sea urchin egg. In: Malacinski, GM., editor. *Developmental Genetics of Higher Organisms. A Primer in Developmental Biology*. MacMillan; New York: 1988. p. 147–170.
- Garcia-Fernandez J, Holland PHW. Archetypal organization of the amphioxus Hox gene cluster. *Nature* 1994;370:563–566. [PubMed: 7914353]
- Grobber K. Die systematische Einteilung des Teirreichs. *Verh Zool-bot Ges Wien* 1908;58:491–511.
- Guigo R, Agarwal P, Abril JF, Burset M, Fickett JW. An assessment of gene prediction accuracy in large DNA sequences. *Genome Res* 2000;10:1631–42. [PubMed: 11042160]
- Halanych KM, Bachele JD, Aguinaldo AMA, Liva SM, Hillis DM, Lake JA. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science* 1995;267:1641–1643. [PubMed: 7886451]
- Hibino T, Loza-Coll MA, Messier C, Majeske A, Cohen A, Terwilliger D, Buckley K, Brockton V, Nair S, Berney K, Fugmann SD, Anderson MK, Pancer Z, Cameron RA, Smith LC, Rast J. The immune gene repertoire encoded in the purple sea urchin genome. *Dev Biol* 2006;300:3439–365.
- Hinegardner R. Cellular DNA content of the Echinodermata. *Comp Biochem Physiol* 1974;49B:219–226.
- Howard-Ashby M, Materna SC, Brown CT, Chen L, Cameron A, Davidson EH. Identification and characterization of homeobox transcription factor genes in *S. purpuratus*, and their expression in embryonic development. *Dev Biol* 2006;300:74–89. [PubMed: 17055477]
- Howard-Ashby M, Brown CT, Materna SC, Chen L, Cameron A, Davidson EH. Gene families encoding transcription factors expressed in early development of *Strongylocentrotus purpuratus*. *Dev Biol* 2006;300:90–107. [PubMed: 17054934]
- InterPro, Consortium. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research* 2001;29:37–40. [PubMed: 11125043]
- Marzluff WF, Sakallah S, Kelkar H. The sea urchin histone gene complement. *Dev Biol* 2006;300:308–320. [PubMed: 17078943]
- Materna SC, Howard-Ashby M, Gray RF, Davidson EH. The C₂H₂ zinc finger genes of *Strongylocentrotus purpuratus* and their expression in embryonic development. *Dev Biol* 2006;300:108–120. [PubMed: 16997293]
- Materna SC, Berney K, Cameron RA. The *Strongylocentrotus purpuratus* genome: A comparative perspective. *Developmental Biology* 2006;300:485–495. [PubMed: 17056028]
- Nair SV, Del Valle H, Gross PS, Terwilliger DP, Smith LC. Macroarray analysis of coelomocyte gene expression in response to LPS in the sea urchin, *Strongylocentrotus purpuratus*. Identification of unexpected immune diversity in an invertebrate. *Physiological Genomics* 2005;22:33–47. [PubMed: 15827237]
- Ohno, S. *Evolution by Gene Duplication*. Springer; New York: 1970.
- Oliveri P, Davidson EH. Gene regulatory network controlling embryonic specification in the sea urchin. *Curr Opin Genet Dev* 2004;14:351–360. [PubMed: 15261650]

- Palumbi SR, Wilson AC. Mitochondrial DNA diversity in the sea urchins *Strongylocentrotus purpuratus* and *S. drobachiensis*. *Evolution* 1987;44:403–415.
- Pederson T. The sea urchin's siren. *Dev Biol* 2006;300:9–14. [PubMed: 17113576]
- Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M. The Ensembl analysis pipeline. *Genome Res* 2004;14:934–941. [PubMed: 15123589]
- Rast JP, Smith LC, Loza-Coll M, Hibino T, Litman GW. Genomic Insights into the Immune System of the Sea Urchin. *Science* 2006;314:952–956. [PubMed: 17095692]
- Robertson AJ, Croce J, Carbonneau S, Voronina E, Miranda E, McClay DR, Coffman JA. The genomic underpinnings of apoptosis in *Strongylocentrotus purpuratus*. *Dev Biol* 2006;300:321–334. [PubMed: 17010332]
- Salamov AA, Solovyev VV. Ab initio gene finding in Drosophila genomic DNA. *Genome Res* 2000;10:516–522. [PubMed: 10779491]
- Samanta MP, Tongprasit W, Istrail S, Cameron RA, Tu Q, Davidson EH, Stolc V. The transcriptome of the sea urchin embryo. *Science* 2006;314:960–962. [PubMed: 17095694]
- Sea Urchin Genome Sequencing Consortium. The Genome of the Sea Urchin *Strongylocentrotus purpuratus*. *Science* 2006;314:941–952. [PubMed: 17095691]
- Sodergren E, Shen Y, Song X, Zhang L, Gibbs RA, Weinstock GM. Shedding genomic light on Aristotle's lantern. *Dev Biol* 2006;300:2–8. [PubMed: 17097628]
- Solovyev, VV. Statistical approaches in Eukaryotic gene prediction. In: Balding, DEA., editor. *Handbook of Statistical Genetics*. John Wiley and Sons, Ltd; 2001. p. 83-127.
- Sonnhammer ELL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: Multiple Sequence Alignments and HMM-Profiles of Protein Domains. *Nucleic Acids Res* 1998;26:320–322. [PubMed: 9399864]
- Souvorov A, Tatusova T, Lipman DJ. Genome annotation with Gnomon—A multistep combined gene prediction tool. *ISMB* 2004:125.
- Stuart JM, Segal E, Koller D, Kim SK. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. 2003;302(5643):249–255.
- Turbeville JM, Schulz JR, Raff RA. Deuterostome phylogeny and the sister group of the chordates - evidence from molecules and morphology. *Mol Biol Evol* 1994;11:648–655. [PubMed: 8078403]
- Wada H, Satoh N. Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18S rDNA. *Proc Natl Acad Sci USA* 1994;91:1801–1804. [PubMed: 8127885]

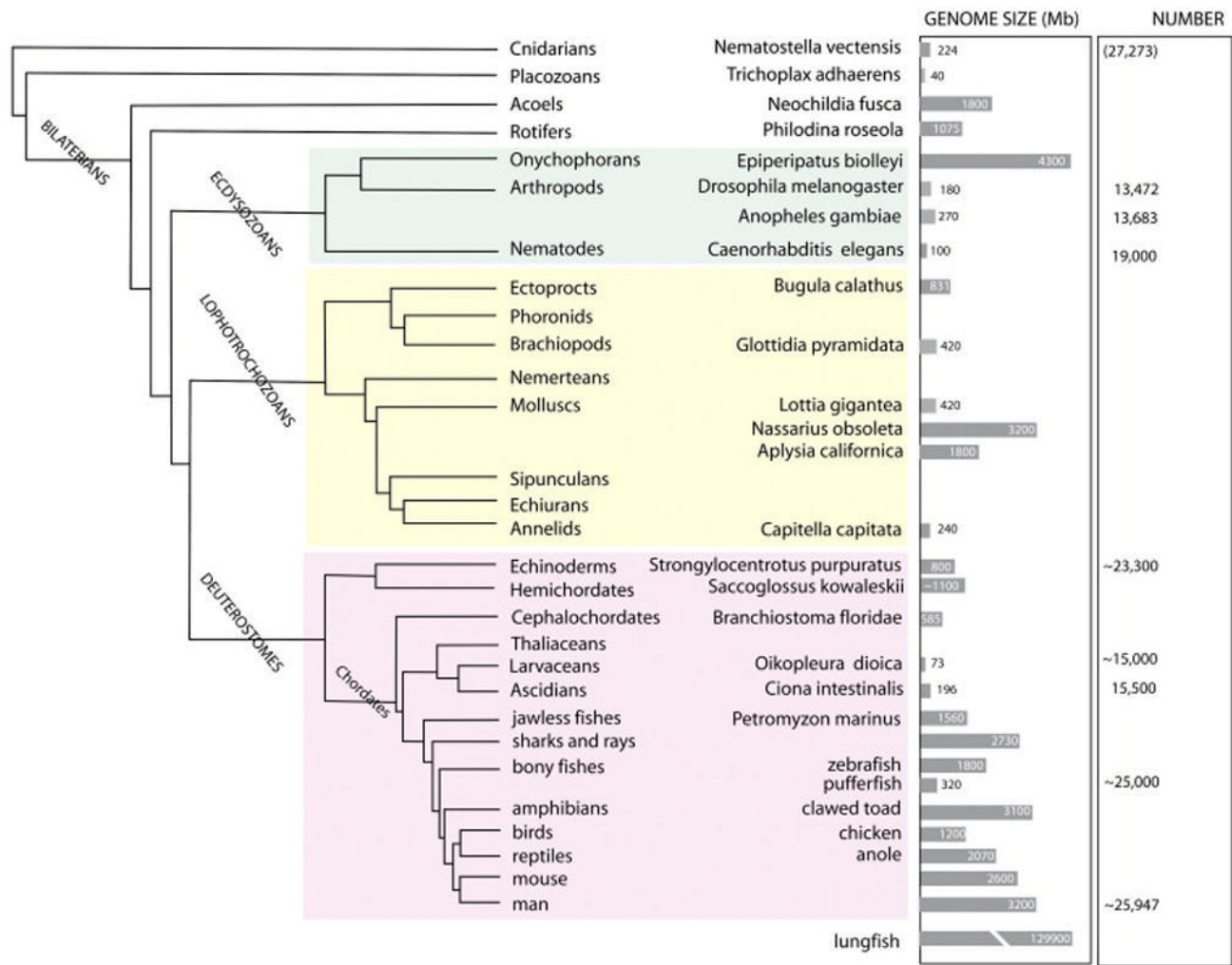
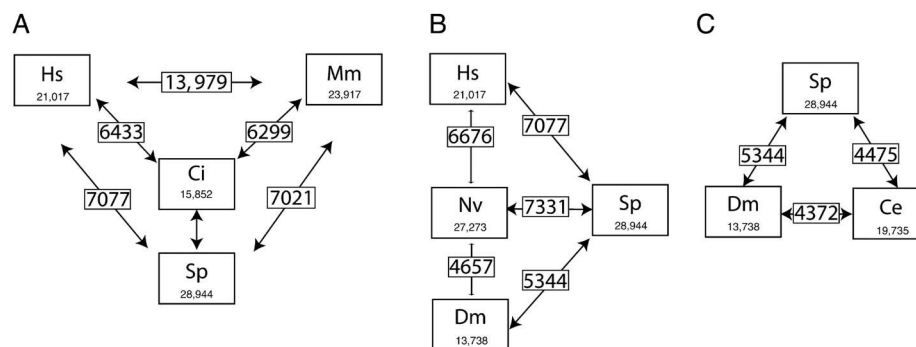


FIGURE 1. A modern phylogeny of selected metazoan animals (see text for references). Individual species in each taxon are indicated along with the genome size and gene number where it is well documented. The deuterostome phyla are shaded pink, the lophotrochozoan, yellow and the ecdysozoans, green. The closest outgroup is thought to be the placozoans but the cnidarians are a better bilaterian outgroup.

**FIGURE 2.**

Gene set comparisons by reciprocal best blast. A. Deuterostome comparison. The species are shown in boxes with the number of reciprocal best blast totals at an upper threshold expectation value of 1×10^{-6} indicated along the line between them. The total number of gene models in the species is indicated within the box. B. Metazoan comparison. C. Comparison between the ecdysozoan genomes and the sea urchin. Abbreviations: Ce, *Caenorhabditis elegans*; Ci, *Ciona intestinalis*; Dm, *Drosophila melanogaster*; Hs, *Homo sapiens*; Mm, *Mus musculus*; Nv, *Nematostella vectensis*; Sp, *Strongylocentrotus purpuratus* (redrawn from Materna et al., 2006).

The classification of gene models by best PFAM hit for the sea urchin and other genomes

ID ^a	Name ^b	S.p. (rank) ^c	M.m. (rank)	C.i. (rank)	D.m. (rank)	C.e.(rank)
PF00001	7tm_1	954 (1)	372 (12)	145 (5)	6 (537)	4 (729)
PF00097	zf-C3HC4	400 (6)	151 (46)	75 (28)	69 (32)	53 (64)
PF00530	SRCR	373 (9)	88 (83)	41 (53)	16 (203)	0 (2430)
PF00125	Histone	335 (10)	20 (393)	6 (384)	11 (310)	1 (1480)
PF00096	zf-C2H2	319 (12)	6 (1063)	1 (1354)	3 (1036)	0 (2518)
PF05729	NACHT	226 (19)	21 (374)	1 (1302)	2 (1321)	3 (991)
PF00059	Lectin_C	218 (20)	4 (1489)	1 (1572)	4 (855)	3 (921)
PF01582	TIR	180 (23)	12 (654)	3 (708)	9 (393)	22 (154)
PF00002	7tm_2	177 (24)	0 (3115)	0 (3115)	0 (3214)	0 (3291)
PF00643	zf-B_box	175 (25)	1 (2453)	0 (2584)	0 (2903)	0 (3093)
PF02494	HYR	154 (27)	3 (1765)	1 (1679)	1 (1925)	1 (1957)

^aIdentification number used in the PFAM database.

^bName given to the domain or motif family in the database.

^cThe value in each category for each species is presented as the total number, with the rank of the total matches in parentheses. Species abbreviations: S.p., *S. purpuratus*; M.m., *M. musculus*; C.i., *C. intestinalis*; D.m., *D. melanogaster*; C.e., *C. elegans*.